Revue-IRS

Revue Internationale de la Recherche Scientifique (Revue-IRS)

ISSN: 2958-8413

Vol. 2, No. 2, Avril 2024

Crainte de l'IA ou méfiance de l'humain envers soi-même ? : Une autopsie de la peur de l'intelligence artificielle

SORO Torna

UFR Sciences de l'Homme et de la Société (SHS), Département de Philosophie

Université Félix HOUPHOUËT BOIGNY (Côte d'Ivoire)

Laboratoire Société, individu, culture (LaSIC)

Résumé : L'intelligence artificielle (IA) ne laisse personne indifférente, universitaire, politique comme citoyen ordinaire. Mais, autant elle suscite de l'espoir, autant elle alimente les inquiétudes. Ses incidences nuisibles sur l'homme sont redoutées. C'est ce qui explique les multiples engagements dans l'élaboration de normes éthiques capables d'encadrer, à la fois la conception des systèmes d'IA et leurs utilisations. Sauf que l'IA ne se conçoit pas toute seule et elle ne s'emploie pas elle-même non plus. L'homme en est le concepteur et l'usager. De ce fait, l'IA ne devrait pas tant faire peur à l'homme. Il apparait opportun de porter un regard attentif sur l'enracinement de la crainte qui alimente ces engagements d'encadrement éthique de l'IA. C'est dans ce sens que ce texte s'est proposé d'examiner l'essence et le fondement de la crainte de l'IA, à travers une approche phénoménologique et critique. Cette analyse révèle que la peur de l'IA est réelle et matériellement fondée, à la fois aux niveaux existentiel et social. Cependant, cette crainte tire sa source de la méfiance de l'homme envers lui-même. Cette méfiance s'alimente, d'une part, de la perspective nuisible des conséquences des égoïsmes naturels de l'être humain et, d'autre part, des insuffisances cognitives intrinsèques et acquises de l'humain lui-même face à l'IA.

Mots clés : Crainte ; Éthique ; Humain ; Intelligence artificielle ; Insuffisances cognitives ; Méfiance.

Abstract: Artificial intelligence (AI) leaves no one indifferent, be they academics, politicians or ordinary citizens. But as much as it inspires hope, it also fuels concern. Its harmful effects on human beings are feared. This explains the many commitments to the development of ethical standards capable of providing a framework for both the design of AI systems and their use. Except that AI does not design itself, nor does it use itself. It is designed and used by humans. As a result, AI should not frighten man so much. It seems appropriate to take a close look at the roots of the fear that fuels these commitments to the ethical framework of AI. It is with this in mind that this text sets out to examine the essence and basis of the fear of AI, using a phenomenological and critical approach. This analysis reveals that the fear of AI is real and materially founded, at both existential and social levels. However, this fear is rooted in man's distrust of himself. This distrust is founded, on the one hand, on the harmful prospect of the consequences of human beings' natural egoisms and, on the other, on the intrinsic and acquired cognitive inadequacies of humans themselves in the face of AI.

Keywords: Fear; Ethics; Human; Artificial intelligence; Cognitive deficiencies; Distrust.

Digital Object Identifier (DOI): https://doi.org/10.5281/zenodo.11099879

1 Introduction

L'avènement de l'Intelligence Artificielle (IA) a mis en perspective de multiples possibilités dans la plupart des domaines de connaissance et d'activité. Elle fait reculer les limites du pouvoir humain, tant du point de vue cognitif que de celui de l'agir humain. Les espoirs sont énormes, avec l'IA. Cependant, autant elle suscite de l'espoir, elle n'engendre pas moins de craintes. Cette peur est manifeste à travers, d'une part, les intentions des concepteurs d'observer un moratoire sur le développement de l'IA (Savornin, 2023 : 1) – ou de trouver un ensemble de normes l'encadrant –, et d'autre part, l'engagement de l'UNESCO (Noiseau et al., 2021 : 1), des comités d'éthique/bioéthique et des universitaires à trouver des canevas d'encadrement éthique de l'IA. Cette recherche de mécanismes d'encadrement éthique de l'IA donne lieu à un nombre important de travaux scientifiques, d'un côté, sur ses bénéfices pour l'homme et, d'un autre, sur les risques auxquels elle expose l'humanité. Cette crainte est légitime, vu que les implications et portées des applications de l'intelligence artificielle sont encore peu déterminées et maîtrisées. Il y a bien d'angles morts dans la gouvernance de l'IA – des aspects encore insoupçonnés auxquels peut s'étendre cette technologie (Farnadi, Alves, Salganik, 2023 : 55). Aussi, la perspective l'introduction de l'IA dans certains domaines, en l'occurrence le domaine militaire, inquiète fortement (Kunertova, 2021 : 3).

Mais, au-delà de l'observabilité de la crainte de l'IA, il peut être intéressant de s'interroger sur l'essence de cette peur. Mieux, il semble nécessaire de chercher à comprendre ce qui est à l'origine des inquiétudes soulevées par l'avènement de l'IA et ses perspectives. Au fond, si l'IA est un artéfact, fruit de l'être humain et destinée à son usage, l'inquiétude qu'elle suscite semble renfermer une réalité inexprimée, ou non cernée. L'humain, auteur et utilisateur de cet artéfact, semble lui-même être mis en cause du point de vue axiologique. C'est pour cela que cette contribution se propose d'analyser l'essence de la crainte suscitée par l'intelligence artificielle, à partir d'une approche phénoménologique et critique. Autrement dit, quel est fondement de la crainte de l'IA chez l'homme ? Ainsi, cette étude procède d'abord par une analyse phénoménale et artéfactuelle de l'IA, puis un examen du fond axiologique de l'humain. Les axes de réflexion retenus sont alors : 1) Analyse phénoménale et prospective de la crainte de l'IA : Des implications éthiques de l'IA ; 2) La crainte de l'IA : une méfiance de l'homme en lui-même.

2 Analyse phénoménale et prospective de la crainte de l'IA : Des implications éthiques de l'IA

Il est question d'analyser, à partir du développement actuel et potentiel de l'IA, la pertinence des craintes liées à l'IA. L'analyse permet de regrouper ces craintes en deux principales catégories, à savoir les craintes existentielles et les craintes sociales. Les craintes existentielles sont celles en lien avec l'être matériel de l'homme, aussi bien dans son être-au-monde présent en tant que sujet que dans son devenir, en tant qu'individu et en tant qu'humain qui reste tel, au sens d'un H. Jonas (2013 : 255). Les craintes sociales, quant à elles, renvoient aux préoccupations en rapport avec les facteurs susceptibles d'altérer la vie en société ou de mettre à mal les acquis sociaux, tant au niveau communautaire qu'individuel.

L'un des premiers aspects, faisant référence à une peur existentielle, est naturellement la menace contre l'intégrité physique de l'humanité. En rapport avec l'intelligence artificielle,

c'est la capacité de cette technologie à mettre en danger l'existence du genre humain ou d'un groupe humain. Il apparait un sentiment d'insécurité, pour la pérennisation de son être-là, face à l'évolution de l'IA. Cette crainte peut être appréhendée à deux niveaux, d'une part dans le potentiel phénoménal de l'IA à nuire et, d'autre part, dans les diverses réactions face à cette technologie.

Au premier niveau, les problèmes de sécurité des systèmes informatiques de l'IA sont à considérer. Ainsi, l'idée d'une vulnérabilité desdits systèmes suscite de la frayeur, au regard de leur pouvoir d'action. Bien plus, vu que ces systèmes ont davantage un fonctionnement autonome, cela fait craindre les revers d'une perte de leur contrôle. C'est pourquoi A. Boily (2020 : 40) indique que la « question de la sécurité des systèmes autonomes est probablement l'une des plus visibles et controversées en éthique de l'IA, puisque son contenu est fréquemment mis de l'avant à titre d'exemple de scénarios potentiellement dystopiques ». Il y a une conscience claire des risques liés à un déficit de protection des systèmes d'IA.

Cette remarque Anne Boily se termine avec la mise en perspective du second niveau d'appréhension de la crainte existentielle suscitée par l'intelligence artificielle. Son expression est caractérisée par une alerte puis des initiatives de protection ou de protestation. Ainsi, face aux projets et perspectives d'association de l'IA à l'accès et la manipulation des armes de destruction massive, diverses alertes ont été enregistrées. Parmi celles-ci, celle de l'OTAN semble la plus significative. En effet, cette organisation attire l'attention du monde sur le potentiel destructeur de la combinaison de l'IA avec d'autres technologies productrices d'armes de destruction à grande échelle – dont l'impact sera inédit. Dans son rapport sur les perspectives technologiques de l'armement dans le monde de 2020 à 2040, l'Organisation de la Science et de la Technologie de l'OTAN relève que

> cet impact se produira principalement grâce à l'utilisation de l'IA intégrée dans d'autres technologies associées telles que la réalité virtuelle/augmentée, l'informatique quantique, l'autonomie, la modélisation et la simulation, l'espace, la recherche sur les matériaux, la fabrication et la logistique, et l'analyse des mégadonnées. L'IA aura des effets transformateurs sur le nucléaire, l'aérospatiale, le cyber, les matériaux et les biotechnologies. […] En outre, une dépendance excessive à l'égard des systèmes d'IA introduira également de nouvelles vulnérabilités importantes et ouvrira la voie à une course aux armements contradictoires en matière d'IA (OTAN, 2020 : 54-55).

L'ampleur réelle de ces transformations reste inconnue à présent, toutefois l'on est conscient qu'elle sera plus dévastatrice que les armes nucléaires dont les traumatismes des effets sur Hiroshima et Nagasaki sont encore présents dans les consciences. Déjà, l'on redoute l'idée d'une guerre nucléaire, en raison de l'improbabilité de la survive de l'humanité aux effets d'une telle aventure. La perspective de produire des armes ayant un potentiel de destruction, devant lequel celui du nucléaire ne serait qu'infime, amène naturellement l'homme à craindre pour son existence, en présence de telles éventualités offertes par l'IA. Cette technologie dans le domaine militaire présente un enjeu existentiel prépondérant pour le genre humain. Cela s'appréhende dans l'importance accordée au changement que l'intelligence artificielle induit dans le domaine de l'armement. Les armes incorporant de l'IA, armes létales autonomes, sont dite de troisième révolution dans l'industrie de l'armement (Russell et al., 2015 : 415).

Au-delà des alertes, les actions entreprises en vue de parer les menaces rattachées à l'intelligence artificielle marquent également ce second niveau de la manifestation de la crainte

existentielle. Ces initiatives sont observables chez les concepteurs de cette technologie et les États. C'est dans ce sens que s'inscrit l'initiative Future of Life Institute (FLI) de Boston. Elle promeut l'engagement volontaire de renoncement à certains développements de l'IA susceptibles de mettre en danger la vie dans un futur proche ou lointain. Dans ce sens, cette initiative incite les acteurs de l'IA et les décideurs politiques à poser des actions concrètes ou à prendre des dispositions règlementaires explicites empêchant l'extension de cette technologie à certains domaines dont celui de l'armement. Dans le cadre de l'initiative FLI, en 2018, « des milliers de chercheurs en intelligence artificielle ont également signé une lettre d'engagement à bannir les « robots tueurs », soit les armes létales autonomes ou « LAWS » (« lethal autonomous weapon systems ») » (Boily, 2020 : 40). Cette action est préventive des menaces que présente l'IA dans ce secteur extrêmement sensible.

Cette attitude précautionniste est également apparente aux niveaux étatique et trans-étatique. Dans le cadre de l'initiative FLI, plusieurs pays, membres des Nations Unies, ont pris à la résolution, en 2018 également, de renoncer aux armes dotées d'IA. Ce type de décision devrait freiner les velléités de course aux armes IA. Au-delà des engagements volontaristes, se précisent des volontés des dispositions contraignantes, par le canal de la législation, surtout dans l'espace de l'Union Européenne. Dans un document d'aide au travail parlementaire, élaboré à l'attention du parlement européen, T. Metzinger (2018 : 34) suggère ceci : « l'Union devrait interdire toute recherche présentant un risque ou ayant pour objectif direct de développer une phénoménologie synthétique sur son territoire, et chercher à conclure des accords internationaux ». La crainte de l'IA n'est, de ce fait, plus simplement spéculative. Elle est saisie et exprimée explicitement par les chercheurs, à l'attention des décideurs et législateurs. La matérialité de cette crainte d'une atteinte à l'existence physique de l'être humain, au-delà de tout optimisme suscité par l'intelligence artificielle, interpelle l'homme quant aux dangers rattachés à son l'emploi dans certains domaines sensibles.

Un autre aspect de la peur existentielle suscitée par l'IA est la probabilité de l'avènement d'une existence marginale de l'humain face à cette technologie. Cette éventualité se rattache, d'une part, au dépassement de l'intelligence humaine par l'intelligence artéfactuelle et, d'autre part, à la perte de la priorité de l'être pensant, de la qualité de référent pour la résolution des difficultés de l'existence. Le dépassement de l'intelligence humain, appréhendé sous le concept de singularité (technologique), est analysé par Ray Kurzweil (2005) dans son ouvrage intitulé The Singularity Is Near. When Humans Transcend Biology. Pour lui, le potentiel de calcul sera pleinement utilisé par les machines intelligentes au tour de 2030. Bien que cela n'est nullement le dépassement de l'intelligence humaine – c'est-à-dire intelligence biologique –, l'on dépasse de loin l'utilisation de l'intelligence biologique dont les capacités sont sous exploitée. En effet, il écrit :

> Cet état des calculs au début des années 2030 ne représentera cependant pas la Singularité, parce que cela ne correspondra pas encore à une profonde expansion de notre intelligence. Au milieu des années 2040, ce millier de dollars de calculs sera égal à 1026 cps, l'intelligence créée par an (pour un coût total de 1012 $) sera donc approximativement un milliard de fois plus puissante que toute l'intelligence humaine aujourd'hui. Cela représentera en effet un changement profond, et c'est pour cette raison que j'ai établi à 2045 la date de la Singularité - qui représente une transformation profonde et perturbatrice des capacités humaines (Kurzweil, 2007 : 149).

À l'évaluation de l'intelligence artificielle, moins de deux décennies après cette analyse Kurzweil, la perspective de la Singularité se profile. Aujourd'hui, l'intelligence artificielle est capable d'exécuter un très large éventail diversifié de taches, avec une célérité inégalable par l'humain (Bubeck et al., 2023 : 92). Selon les chercheurs de Microsoft Research, les formes actuelles de l'IA dépassent déjà l'intelligence de l'être humain. Leur rapport de mars 2023 sur les tests effectués sur GPT-4, « démontre que dans toutes les tâches mesurées, les performances de GPT-4 sont proches ou supérieures aux performances humaines » (D. Anctil, 2023 : 69). Ainsi, l'humain se fait progressivement dépasser par ses artéfacts en termes d'intelligence.

Mais, l'inquiétude n'est pas simplement le dépassement de l'intelligence humaine, mais plutôt ses implications éthiques pour l'homme. Car, les perspectives d'une mise en marge de l'être humain, dans ce qui est considéré comme sa marque distinctive – l'intelligence, la pensée –, constituent l'une des véritables raisons de la crainte éprouvée face à l'IA. Il y a comme une dépréciation de la substance de l'humain. C'est pourquoi Marosan (2019 : 148) estime que, l'homme, d'une « manière générale, en se robotisant, […] finit forcément par se déshumaniser. D'une manière précise, il profite du développement de l'IA sur le plan de l'utilité, de l'efficacité, et du rendement, tout en perdant jusqu'à un certain point des éléments qui le composent dans ce qui fait de lui un être à proprement parler imparfait, et donc tout le contraire d'un robot ». Ainsi, l'IA dépouille l'homme de ses prérogatives du fournisseur central de la connaissance et le met dans la posture d'un piètre penseur imparfait et fortement limité, en comparaison à ce que fournit l'artéfact intelligent. Il y a, ici, une perte de contrôle sur la pensée qui se profile et, bien plus, une perte de contrôle des systèmes IA, au regard de leur caractère extrêmement dynamique et imprévisible (Lahoual & Fréjus, 2018 : 418) – même si certains analystes estiment que ces systèmes sont encore loin de ce stade d'autonomie absolue (Kunertova, 2021 : 1). Tout cela ne fait que renforcer la peur de l'homme face à l'IA, quant à son existence et au sens de celle-ci – si l'on considère la pertinence du cogito cartésien.

À ces craintes existentielles se superposent des craintes sociales qui, en dans le fond, ne sont que le prolongement des premières. Les craintes d'ordre social sont d'abord en lien avec la liberté des personnes. L'IA donne des possibilités inestimables de surveillance et de contrôle. Arrimée aux systèmes de vidéosurveillance, elle optimise l'identification et le suivi constant des individus. Ses systèmes raffinent le contrôle des habitudes et étend, par-là, les questions de l'autonomie des individus dans leurs simples choix. En effet, ce que l'on peut appeler identification-fichage des individus favorise l'influence des décisions par le traçage des activités sur le net, et des lieux fréquentés, puis la proposition constante de produits (Lahoual & Fréjus, 2018 : 419 ; COMEST, 2017 : 12, 56 ; Voarino & Régis, 2023 : 99-100). Cette influence peut s'étendre aux choix politiques. C'est alors l'avenir des systèmes politiques qui est en jeu – au regard des perspectives de manipulation des décisions des individus. Par ces questions de liberté personnelle et politique, l'IA semble donner à la biopolitique ses heures de gloire.

La crainte sociale procède aussi de la perspective de dislocation du lien social entre humains, et avec lui le sens de la vie et des relations intersubjectives. C'est comme si l'IA mène l'homme vers horizon social indéterminé, mais où il est certain que la société future risque d'être régit par de l'inter-artéfactuel (Marosan, 2019 : 147). Les relations entre humains médient par l'artéfact intelligent – qui pourrait transmettre une commission, assurer une conversation, faire

un retour de commission, accomplir une mission et en faire un rapport précis, prendre des décisions au nom de celui pour lequel elle agit, etc. (ce qui ne requiert qu'une fusion de l'IA avec la robotique ou une simple intégration). Cette perspective sociale incertaine, et radicalement différente de celle connue jusqu'à présent, suscite naturellement de la crainte. Cette inquiétude s'amplifie lorsque l'on considère la perspective d'une aggravation des problèmes sociaux actuels, en l'occurrence la discrimination, le racisme et les inégalités sociales (Ravet, 2018 : 5 ; Häggström, 2018 : 24 ; Commission Européenne Pour l'Efficacité de la Justice (CEPEJ), 2019 : 58).

L'un des aspects de cette crainte sociale est le devenir de l'éducation et de la connaissance, et par ricochet de la pensée. En effet, avec l'intelligence artificielle, le recours à la réflexion humaine pour résoudre les problèmes scolaires tend à diminuer et l'on assiste à l'expansion d'une utilisation systématiques des outils logiciels pour traiter les devoirs (Anctil, 2023 : 71). De plus, Anctil (2023 : 71) montre que, « contrairement à une idée répandue, il est pratiquement impossible de détecter l'utilisation frauduleuse des nouvelles applications dans un travail scolaire. La raison est simple : le contenu généré par l'IA est un contenu original comparable à celui d'un rédacteur expert qui s'inspire de ses sources ». C'est donc la construction du savoir qui est en cause. L'avenir de la science, telle que connue, est en jeu. L'interrogation, ici, est de savoir si l'IA peut assurer la construction future des connaissances. Puisque les connaissances sont élaborées dans la recherche de solution à des problèmes rencontrés par l'être humain dans existence, si l'on considère que l'IA assure la relève de la pensée, le savoir construit par cette technologie sera au service de qui ? Et répondra-t-il aux besoins existentiels de l'humain ? Si malencontreusement l'humanité s'abrutit du point de vue connaissance, l'IA subsistera-t-elle ? Au fond, c'est également l'avenir de l'intelligence artificielle est qui est en jeu, si la production de la connaissance tombe dans une impasse. Par ailleurs, cela introduit un troisième type de crainte, à savoir celle de l'intégrité de l'IA. Se nourrissant et s'améliorant, par le deep learning, à partir des bases de données disponibles en ligne, l'IA dépend des informations disponibles. De ce fait, la qualité des rendus de cette technologie est tributaire de la qualité des informations mises en ligne. Ainsi, l'abrutissement de l'humain engendrerait la naissance d'une stupidité artificielle en lieu et place de l'intelligence artificielle.

Il apparait de l'analyse ci-dessus que la crainte de l'IA se fonde à la fois aux niveaux existentiel et social. Mais, l'observation à faire est que cette technologie est, avant tout, une production humaine et est à l'usage de l'humain. C'est lui qui l'alimente et l'emploie à ses propres fins – jusqu'à présent. C'est pourquoi l'homme pourrait légitiment s'interroger sur les fondements de la crainte éprouvée par l'être humain face à sa création et son outil qu'est l'IA. Ne devrait-on pas quêter dans le sens de la confiance que l'homme a en lui-même ? Ne se méfierait-il pas de soi-même ?

3 De la crainte de l'IA : une méfiance de l'homme envers soi-même

La crainte est, au fond, l'expression d'une détresse face à une réalité dont la maîtrise et les effets que devraient subir le craintif sont incertains. Cette détresse est un appel à la rescousse – un SOS –, afin de pallier les facteurs de détresse ou tout au moins de les amoindrir. Ainsi, la crainte est un appel à une éthique en vue, ici, de contenir les menaces des intelligences artificielles susceptibles de porter un préjudice à l'humanité. Mais, ici, l'humain est la source originelle de

l'action avec l'IA, et il est aussi celui qui est menacé par l'IA – ou qui se sent menacé –, et il est également celui à qui l'appel à l'aide est adressé. Il ne devrait pas y avoir de crainte, puisque l'humain se retrouve aux quatre niveaux de la boucle – concepteur, utilisateur, menacé et secouriste. L'acteur est le craintif. Or, vu la réalité, à la fois au niveau des universitaires, des politiques et des organismes transnationaux, la crainte est omniprésente. Il semble y avoir un fond inexprimé ou indiscerné de cette crainte.

Si l'homme est inquiet devant sa créature que lui-même utilise, l'on peut par analyse logique relever qu'il se craint lui-même. Se craindre, c'est avoir peur de subir des préjudices de sa propre part, et ce soit intentionnellement, soit par mégarde. Cette crainte de soi est en réalité une crise de la confiance en soi – ici, de l'être humain en soi-même. Cette crise de la confiance de l'homme en lui-même est aussi l'expression d'une méfiance envers soi-même. L'enracinement de la crainte de l'intelligence artificielle dans la méfiance de l'humain envers soi-même se fonde, si l'on analyse certains aspects caractéristiques de l'humain, dans son déploiement existentiel.

L'un des premiers fondements de cette méfiance de l'homme envers soi-même est son potentiel de nuisibilité pour lui-même. Indépendamment d'une technologie au potentiel destructeur, dotée d'une IA, l'humanité est déjà à mesure de porter un important préjudice à son existence ou à son intégrité physique. Les moyens pour le faire ne manquent pas. Et, cela est évident depuis les deux guerres mondiales, tout au moins à la seconde qui a révélé le niveau d'horreur dont est capable le genre humain, par le biais son nouveau pouvoir technoscientifique d'action. C'est dans ce sens que Brundage (2018 : 20) affirme que « même sans IA plus avancée, l'être humain possède (au moins depuis la mise au point d'armes nucléaires) la capacité de se détruire lui-même ». La conscience de cette réalité et les possibilités d'une action irréversible président les ambitions et initiatives – même si celles-ci restent limitées – de dénucléarisation de la planète. Devant ce potentiel, son amplification avec les systèmes de l'IA fait davantage redouter le niveau de nuisibilité.

Mais, l'une des plus importantes raisons de la méfiance de l'homme envers soi-même est l'existence de tendances au suicide. L'existence de cette réalité inhérente à l'être-au-monde de l'humain constitue une source probable de la destruction de son genre, au regard des facilités offertes par les outils de l'IA. Ici, c'est l'éventualité qu'un individu risque l'être de l'humanité qui engendre la méfiance envers l'homme. Les intentions suicidaires d'un seul peuvent compromettre l'existence de l'humanité dans son entièreté. Il est certain qu'il est peu probable que l'on aboutisse à un accord de l'espèce humaine – unanime, ni même majoritaire – de son autodestruction ou l'exécution du testament d'un Arne Naess ou de la deep ecology sur la réduction drastique de la population humaine (Naess, 1990 : 31-32). Ce qui semble correspondre à ce que relève Durkheim sur la distinction entre les consciences individuelles, notamment sur le suicide, et l'orientation globale de la société. Pour lui, « *de toutes les consciences particulières qui composent la grande masse de la nation, il n'en est aucune par rapport à laquelle le courant collectif ne soit extérieur presque en totalité, puisque chacune d'elles n'en contient qu'une parcelle* » (Durkheim, 1897 : 357). Ainsi, l'intention du suicidaire n'est jamais partagée par toutes les consciences individuelles, et elle ne correspond guère à la volonté collective. Toute initiative, quels qu'en soient les fondements théoriques ou idéologiques, ne s'aurait traduire l'assentiment de l'humanité, laquelle s'incarne en chaque

individu et qu'aucun ne résume et ne devrait compromettre la continuation de son être-authentique-au-monde – comme le rappelle Jonas (2013 : 195).

Pourtant, le pouvoir dont se dote le genre humain met son gite, la Terre, dans la posture d'une poudrière dans un camp. Nul besoin de l'accord de tout le camp ou d'un supérieur, pour y mettre le feu et mettre en danger tous les occupants du camp – l'humanité, un pays, un continent, une communauté, selon l'étendu de l'onde de choc. Toute personne y ayant accès peut le faire. Or, l'IA facilite à l'homme l'accès aux ressources incendiaires de la poudrière- Terre. La possibilité de l'accès de personnes suicidaires à ces ressources alimente la crainte de l'homme quant à une vulgarisation de l'IA dans tous les domaines. L'être humain n'a pas de confiance absolue en ses propres capacités et intentions à assumer et assurer sa survie continue. Cette méfiance de l'homme envers soi-même vient de la peur des initiatives suicidaires et périlleuses pour tous.

Un deuxième fondement de la méfiance de l'homme envers soi est la perspective de la négligence humaine ou d'un déficit de responsabilité dans la manipulation des systèmes d'intelligence artificielle. Il n'y a aucune garantie que tous utilisent l'IA avec un sens de responsabilité qui soit à la hauteur du danger contenu dans ce que l'on pourrait appeler le risque IA – c'est-à-dire la nuisibilité potentielle rattachée aux usages de l'intelligence artificielle. Cela est dotant plus probable que l'IA ou les systèmes l'intégrant ne sont pas une conception centralisée dont les propriétés sont connues et susceptibles d'être inspectées auprès des utilisateurs, la question la responsabilité devient cruciale et préoccupe. Toute personne ou structure disposant des ressources techniques et/ou financières susceptibles de concevoir un système d'IA peut en élaborer, selon ses objectifs et besoins, avec les propriétés et caractéristiques désirées. Cependant, les avantages et inconvénients de ces systèmes n'en sont pas pour autant réduits. Ils peuvent être, même, plus importants. Une négligence dans l'emploi serait aussi préjudiciable que tout autre système d'IA. C'est d'ailleurs l'inquiétude de la Commission Européenne Pour l'Efficacité de la Justice (CEPEJ) (2019 : 57) qui indique que

> l'opacité des processus de fonctionnement des algorithmes conçus par les entreprises privées (qui revendiquent leur propriété intellectuelle) a été une autre source d'inquiétude. Si l'on tient compte du fait qu'elles tiennent leurs données sources des autorités étatiques elles-mêmes, leur absence d'esprit de responsabilité, quant aux règles relatives à la protection des données (accountability) vis-à-vis des citoyens pose un problème démocratique majeur.

Un usage négligeant, pour la CEPEJ, est appréhendé comme une menace pour les systèmes politiques, particulièrement la démocratie, au regard des incidences que cela peut avoir sur les individus, en termes de liberté, de choix ou d'initiatives personnelles ou collectives. Les balises éthiques ou législatives envisagées sur l'IA visent justement à prévenir les négligences dans l'usage des systèmes intégrant cette technologie et à engager la responsabilité des concepteurs et utilisateurs. Car, la négligence ou le déficit d'usage responsable est bien réel dans l'utilisation des outils techniques ou informatiques. L'étude de Laura Ellyson (2018) sur la responsabilité criminelle dans le cas de l'IA – tant pour l'utilisateur que le concepteur – met en évidence les possibilités d'établir juridiquement la négligence. Elle montre que la négligence peut être criminelle ou pénale (Ellyson, 2018 : 888-889). Le fondement de l'éventualité de la négligence fait craindre les effets que cela pourrait avoir, avec un outil dont l'étendue des revers reste indéterminée. C'est, de ce fait, la capacité de l'être humain à utiliser l'IA avec responsabilité qui est mise en doute. L'homme se méfie alors de soi-même, devant l'intelligence artificielle.

La prévalence de l'égoïsme sur le bien de l'humanité et l'imprévisibilité des intentions de l'homme participent également au renforcement de la méfiance de l'homme envers soi-même. Parmi les caractéristiques fondamentales de l'humain, fait partie son égoïsme. Chez Hobbes, par exemple, cette caractéristique est dominante dans la définition de l'homme. Plus précisément, Le Citoyen de Hobbes met l'accent sur la méchanceté de l'être humain (Hobbes, 1982 : 73, 83). Pour lui, n'étant pas méchant originellement, un ensemble de dispositions le rendent ainsi. « Parmi ces dispositions, l'égoïsme est la plus fondamentale. L'homme naturel de Hobbes est foncièrement égoïste. Ce qu'il recherche d'abord et avant tout, c'est son bien-être ; ce qu'il a en vue dans ses actions, c'est toujours son profit ou sa "fierté" » (Nguyen, 1986 : 257). En d'autres termes, l'égoïsme domine les faits et décisions de l'homme. Dans la même perspective, Adam Smith estime que c'est l'égoïsme des individus, la poursuite de leurs intérêts propres, qui assure le fonctionnement de l'économie.

En présence du pouvoir ambivalent de l'IA, cet égoïsme originel constitue, en lui-même, une menace, et par ricochet un autre élément qui aiguise la méfiance de l'homme à son propre égard. Une utilisation exclusivement égoïste de ce pouvoir ne rechercherait que les intérêts de l'utilisateur, tout en faisant fi des répercutions sur ensemble de la communauté humaine ou sur autrui. Bien plus, des biais algorithmiques des systèmes IA peuvent provenir d'intentions égoïstes. Or, il est difficile de garantir un contrôle des conceptions à la base. La connaissance de l'existence d'un système IA se fait au moment où il est déjà beaucoup avancé ou est soumis à l'usage. Ainsi, des biais volontaires peuvent y être introduits, selon les intentions du concepteur ou de son financier. Ces biais sont possibles dans l'écriture des algorithmes – la programmation – ou dans les données utilisées pour l'apprentissage des systèmes d'IA (Maclure et Saint-Pierre, 2018 : 754). Ici, l'inquiétude se situe au niveau du choix délibéré de faire fi des risques – dans de la conception et/ou dans l'utilisation de l'IA – pour les autres, du fait de la poursuite d'un profit personnel.

La crainte des biais volontaires remet en question la probité du concepteur et la peur des usages exclusivement égoïstes, celle de l'utilisateur. Dans les deux cas, il y a une crise de confiance en l'être humain, et par conséquent une méfiance quant à son intégrité morale. Le problème, dans ce cas, n'est pas l'IA, mais plutôt l'homme lui-même, concepteur et l'utilisateur, ce qu'il peut décider de faire de son artéfact intelligent. La crainte de l'IA, c'est alors un problème de confiance en l'homme et non la technologie elle-même.

Enfin, le dernier aspect fondant la méfiance de l'être humain en soi-même, que l'on peut mettre en évidence, est son insuffisance cognitive. Ici, il ne s'agit nullement d'une crainte liée à ce que l'homme peut faire de compromettant pour lui-même. Il est plutôt question de ses insuffisances, en particulier cognitives. La crainte est de ne pas pouvoir égaler ou dépasser l'artéfact dans l'intelligence. Il y a de la méfiance à l'égard des capacités cognitives de l'humain. Cette insuffisance cognitive peut émaner de deux sources : elle peut être intrinsèque ou acquise. L'on peut parler ainsi d'une insuffisance cognitive intrinsèque de l'humain (ICIH) – ou insuffisance cognitive naturelle de l'humain – et d'une insuffisance cognitive acquise de l'humain (ICAH). L'insuffisance cognitive intrinsèque de l'humain (ICIH) est due au fait que la nature, la constitution biologique de l'homme, ne lui permet pas d'atteindre le niveau d'intelligence de l'IA, c'est-à-dire de mener une réflexion lui permettant – peu importe le temps qu'on lui accorderait – de résoudre des difficultés intellectuelles que peut résoudre l'artéfact, si celui-ci

ne se fait assister par d'autres artéfacts. Cette insuffisance lui semble inhérente. Car, comme l'indique Saint-Affrique (2022 : 5), au-delà de la fatigabilité, le cerveau est beaucoup limité dans sa capacité de traiter les informations, si l'on doit le mettre en rapport avec un ordinateur. Aussi, l'ICIH est mise en évidence par, non pas une incapacité cognitive d'atteindre le niveau d'intelligence de l'IA, mais le fait que la complexité des procédés rend la réponse quasi inaccessible par l'intelligence biologique humaine, sans assistance d'outils artificiels – cette ICIH est de degré moindre que la précédente. Cette éventualité qui est déjà perceptible alimente une forte inquiétude envers les artéfacts intelligents (Maclure et Saint-Pierre, 2018 : 749, 750). Cette crainte de l'IA exprime celle de l'ICIH et une interrogation des capacités intellectuelles de l'homme. Ce questionnement de sa cognition n'est rien d'autre que l'expression d'une méfiance de l'homme vis-à-vis de lui-même, en ce qui concerne son aptitude à garder sa position de penseur et de maître de la pensée pensante.

L'insuffisance cognitive acquise de l'humain (ICAH), quant à elle, est tributaire de la paresse, du renoncement à la culture ou des troubles de l'apprentissage. Ces facteurs sont l'émanation de l'omniprésence de l'artéfact dans l'univers de la pensée de l'homme (Romero, 2018). En effet, la facilité offerte par l'IA pour traiter quasiment toutes les difficultés intellectuelles est susceptible d'altérer l'activité cérébrale et de plonger le cerveau dans une sorte de léthargie cognitive ou d'hibernation qui l'atrophie. Aussi les troubles de la concentration sont à même de produire le même résultat – l'ICAH. Ces troubles sont déjà perceptibles dans le rendement des apprenants, vu que l'hyper-connectivité a une incidence sur le niveau et la qualité de l'apprentissage. Ce qui dégrade déjà les niveaux élémentaires de la connaissance, à savoir la construction de raisonnements, de discours, d'argumentations. Aujourd'hui, « l'argumentation exhaustive et soutenue n'est plus valorisée car seule la #punchline bien placée compte ; […] la vie se résume de plus en plus à une série d'anecdotes qu'il faut impérativement exposer sous la forme d'un #thread à nos abonnés » (Marosan, 2019 : 147). Par accumulation générationnelle, la cognition humaine s'étiole et laisse à l'IA tout le champ de la réflexion. L'avènement de l'ICAH sera comme une résignation face à une intelligence artificielle ou un abandon de son sort à celle-ci. Que le genre humain laisse son devenir aux soins de l'IA, c'est bien une raison d'inquiétude. C'est pourquoi l'ICIH, comme l'ICAH, se présente comme un facteur de base de la méfiance de l'homme envers soi-même.

La perspective de la manifestation de l'ICIH et de l'ICAH place alors l'homme dans un état de malaise et met en exergue sa méfiance envers soi-même. La recherche d'une éthique vise à élaborer un cadre de confiance. Car ce sont, en réalité, les implications éthiques des insuffisances éthiques qui alimentent la peur à l'égard de l'IA. Cette crainte provient de la méfiance envers l'homme, étant donné que les conséquences de l'ICIH et l'ICAH sont encore indéterminées pour l'avenir de l'homme, son intégrité physique et morale. Il ressort, de ce fait, de cette analyse prospective et critique que la crainte de l'IA procède de la méfiance de l'être humain envers soi-même. Cette méfiance se fonde sur une diversité de caractéristiques qui lui sont inhérentes. L'élaboration d'une éthique de l'IA se doit de tenir compte de cette méfiance de l'homme envers lui-même.

4 Conclusion

En définitive, la peur de l'IA est réelle et peut être appréhendée essentiellement en deux catégories, à savoir la crainte existentielle et la crainte sociale. La première provient des menaces existentielles, c'est-à-dire celles qui sont susceptibles de compromettre l'existence et l'intégrité physiques de l'être humain. Elle est aussi en lien avec les risques d'exposition de l'humanité à une existence marginale face à l'IA, notamment avec l'avènement de la singularité technologique. La seconde catégorie de peur, la crainte d'ordre social, est liée aux facteurs IA capables d'altérer le lien social ou les conditions de vie en société. Elle est engendrée par les perspectives d'atteinte aux libertés individuelles et politiques de l'homme, de disparition des relations sociales intersubjectives et la problématique de l'avenir de l'éducation et de la connaissance.

L'autopsie de la crainte de l'IA, l'examen des causes de la peur de l'intelligence artificielle, permet de découvrir qu'elle procède d'une méfiance de l'être humain envers soi-même. Cette méfiance se fonde sur quatre (4) éléments caractéristiques de l'homme, à savoir son potentiel naturel de nuisibilité pour lui-même, la négligence et l'imperfection humaine dans la conception et l'utilisation des systèmes d'IA, l'égoïsme originel de l'être humain et les insuffisances cognitives de l'humain (ICIH et ICAH) face à l'IA. La recherche d'une éthique de l'intelligence artificielle a, justement, pour finalité d'élaborer un cadre normatif. Ce cadre axiologique est appelé à servir de référence en matière d'IA, là où certaines caractéristiques de l'homme engendrent chez lui de la méfiance à son propre égard. De ce fait, une éthique de l'IA se doit de tenir compte de cette méfiance de l'homme envers soi-même, et d'envisager la restauration de la confiance.

REFERENCES

[1] ALLIANCE DES SCIENCES ET TECHNOLOGIES DU NUMÉRIQUE – ALLISTENE. 2014. *Éthique de la recherche en robotique*. Rapport n° 1 de la CERNA, Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene.

[2] ANCTIL, D. 2023. « L'éducation supérieure à l'ère de l'IA générative », *Pédagogie collégiale,* 36(3) : 66-71.

[3] BOILY, A., 2020. *Tensions en éthique de l'intelligence artificielle (IA) : Un guide herméneutique pour les décideurs politiques*, Université de Montréal, Décembre 2020.

[4] BRUNDAGE, M. 2018. « Montée en puissance de l'humanité : les raisons d'un optimisme conditionnel à l'égard de l'intelligence artificielle », in BENTLEY, P. J., BRUNDAGE, M., HÄGGSTRÖM, O., METZINGER T. *Faut-il craindre l'avenir de l'intelligence artificielle ? : Analyse approfondie*. Bruxelles : Union européenne-STOA (PE 614.547) : 16-22.

[5] BUBECK, S. et al. 2023. "Sparks of Artificial General Intelligence. Early experiments with GPT-4", *arXiv preprint arXiv:2303.12712*.

[6] COMMISSION EUROPÉENNE POUR L'EFFICACITÉ DE LA JUSTICE (CEPEJ), 2019. *Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement*, Adoptée lors de la 31e réunion plénière de la CEPEJ (Strasbourg, 3-4 décembre 2018), Strasbourg : Conseil de l'Europe.

[7] COMMISSION MONDIALE D'ÉTHIQUE DES CONNAISSANCES SCIENTIFIQUES ET DES TECHNOLOGIES – COMEST, 2017. *Rapport de la COMEST sur l'éthique de la robotique*, Paris, COMEST- SHS/YES/COMEST-10/17/2 REV.

[8] DURKHEIM, E., 1897. *Le Suicide, étude de sociologie*, Paris : Ancienne Librairie Germer Baillière et C^{ie}.

[9] ELLYSON L. 2018. « La responsabilité criminelle et l'intelligence artificielle : quelques pistes de réflexion », *Les Cahiers de propriété intellectuelle*, 30(3) : 879-893.

[10] FARNADI, G., ALVES, A. L. D. L., SALGANIK, R. 2023. « Le secteur de l'IA du point de vue de l'éthique », in Prud'homme, B., Régis, C. & Farnadi, G. (eds). *Angles morts de la gouvernance de l'IA*, Paris/Montréal : UNESCO/Mila-Institut québécois d'intelligence artificielle : 31-56.

[11] HÄGGSTRÖM, O. 2018. « Observations sur l'intelligence artificielle et l'optimisme rationnel », in BENTLEY, P. J., BRUNDAGE, M., HÄGGSTRÖM, O., METZINGER T. *Faut-il craindre l'avenir de l'intelligence artificielle ? : Analyse approfondie*, Bruxelles, Union européenne-STOA (PE 614.547) : 23-31.

[12] HOBBES, T. 1982, *Le Citoyen*, trad. Samuel Sorbière, Paris : Flammarion.

[13] JONAS Hans, 2013. *Le principe responsabilité : Une éthique pour la civilisation technologique*, trad. Jean Greisch, Paris : Flammarion.

[14] KUNERTOVA, D. 2021, « Les robots militaires : la réalité rejoint la fiction », *Center for Security Studies (CSS)*, ETH Zürich, No 292 : 1-4.

[15] KURZWEIL, R. 2005. *The Singularity Is Near. When Humans Transcend Biology*, New York: Viking Penguin.

[16] KURZWEIL, R. 2007. *Humanité 2.0 : La Bible du changement*, trad. Adeline Mesmin, Paris : M21.

[17] LAFONTAINE, C. 2020. « Bio-objets : enjeux et perspectives de la civilisation *in vitro* », *Revue Médecine et Philosophie*, (4) – 2020 : pp. 31-33.

[18] LAHOUAL, D., Fréjus, M. 2018. « Conception d'interactions éthiques et durables entre l'humain et les systèmes d'intelligence artificielle : Le cas de l'expérience vécue des usagers de l'IA vocale », *Revue d'intelligence artificielle (RIA)*– n° 4/2018 : 417-445.

[19] MACLURE, J. & SAINT-PIERRE, M.-N. 2018. « Le nouvel âge de l'intelligence artificielle : une synthèse des enjeux éthiques », *Les Cahiers de propriété intellectuelle*, 30(3) : 741-765.

[20] MAROSAN, M. I. 2019. « Le devenir robot de l'humain », *Argument*, 21(2) :142-148.

[21] MÉNISSIER, T. 2019. « Quelle éthique pour l'IA ? : Naissance et développements de l'intelligence artificielle à Grenoble », Académie Delphinale, Grenoble : France.

[22] METZINGER, T. 2018. « Vers une charte mondiale de l'intelligence artificielle », in BENTLEY Peter J., BRUNDAGE Miles, HÄGGSTRÖM Olle, METZINGER Thomas, *Faut-il craindre l'avenir de l'intelligence artificielle ? : Analyse approfondie*, Bruxelles : Union européenne-STOA (PE 614.547) : 32-39.

[23] NAESS, A. 1990. *Ecology, community and lifestyle: outline of an ecosophy*, Cambridge: Cambridge University Press.

[24] NATO SCIENCE & TECHNOLOGY ORGANIZATION. 2020. *Science & Technology Trends 2020-2040: Exploring the S&T Edge*, Brussels: NATO Science & Technology Organization.

[25] NGUYEN, V.-D. 1986. « La critique des anthropologies et le *Discours sur l'inégalité* de J.J. Rousseau », *Philosophiques*, 13(2) : 253-266, https://doi.org/10.7202/203319ar (Consulté le 30/12/2023).

[26] NOISEAU, P. et al. 2021. « Le dialogue inclusif sur l'éthique de l'IA : délibération en ligne citoyenne et internationale pour l'UNESCO », *Communication, technologies et développement* [En ligne], n° 10 : 1-13, https://journals.openedition.org/ctd/4294 (Consulté le 02/01/2024).

[27] NORLAIN, B. (dir.). 2021. *Les nouvelles technologies et la stratégie nucléaire.* Paris : Initiatives pour le désarmement nucléaire (IDN).

[28] RAVET, J.-C. 2018. « L'éthique et l'intelligence artificielle », *Relations*, n° 795 : 5-5.

[29] ROMERO, M. 2018. « Intelligence artificielle et pensée humaine », *EpiNet* **205** (2018), https://www.epi.asso.fr/revue/articles/a1805c.htm (Consulté le 02/01/2024).

[30] RUSSELL, S., HAUERT, S., ALTMAN, R. et VELOSCO, M. 2015. "Robotics: Ethics of artificial intelligence", *Nature*, 521(7553): 415-418. doi: 10.1038/521415a. PMID: 26017428.

[31] SAINT-AFFRIQUE, D. D. 2022. « Intelligence artificielle et médecine : quelles règles éthiques et juridiques pour une IA responsable ? », *Médecine & Droit*, 2022(172) : 5-7. SAVORNIN, G. 2023. « Un moratoire pour l'IA ? », *Face au risque*, n° 592 : 1-1.

[32] VIAL, A. 2022. *Systèmes d'intelligence artificielle et responsabilité civile, Droit positif et proposition de réforme*, Thèse présentée et soutenue à Besançon, le 12 décembre 2022.

[33] VOARINO, N. & RÉGIS, C. 2023. « Les dilemmes dans l'angle mort du développement responsable de l'intelligence artificielle en temps de pandémie », in Prud'homme, B., Régis, C. & Farnadi, G. (eds). *Angles morts de la gouvernance de l'IA*. Paris/Montréal : UNESCO/Mila-Institut québécois d'intelligence artificielle : 95-116.